Can extensive reticulation and concerted evolution result in a cladistically structured molecular data set?

Gonzalo Nieto Feliner[1], Javier Fuertes Aguilar[1], and Josep A. Rosselló[2]

[1]Real Jardín Botánico, CSIC. Plaza de Murillo, 2, 28014 Madrid, Spain.

[2]Jardí Botànic, Universitat de Valencia, c/ Quart 80, 46008 Valencia, Spain.

Running title: “Structure in a reticulated and concerted data set”

ABSTRACT

Hierarchy is the main criterion for informativeness in a data set, even if no explicit reference to evolution as a causal process is provided. Sequence data (nuclear ribosomal DNA ITS) from *Armeria* (Plumbaginaceae) contains a certain amount of hierarchical structure as suggested by data decisiveness (DD) and distribution of tree lengths (DTL). However, ancillary evidence suggests that extensive gene flow and biased concerted evolution in these multi-copy regions have significantly shaped the ITS data set. This argument is discussed using parsimony analysis of four data sets, constructed by combining wild sequences with those from different generations of artificial hybrids (wild + F_1, F_2, and backcrosses; wild + backcrosses; wild + F_1; wild + F_2). As compared to the F_1 hybrids, F_2 show a certain degree of homogenization in polymorphic sites. This effect reduces topological disruption caused by F_1 and is considered to be illustrative of how extensive gene flow and biased concerted evolution may have modeled the wild ITS data. The possibility that hierarchy has arisen as a result of —or despite a significant contribution from— those two such potentially perturbing forces raises the question of what kind of signal are we recovering from this molecular data set.

Table of Contents

INTRODUCTION

The mildest impact that reticulation may have on phylogenetic reconstruction is raising the uncertainty in the results. Whether the data set consists of morphological (Humphries, 1983; Funk,1985; McDade, 1995; Skála and Zrzavý,1994) or molecular characters (Rieseberg and Soltis, 1991; Rieseberg and Ellstrand, 1993; Doyle, 1992), two fundamental problems must be faced: the disruption of the hierarchical patterns of character distribution which parsimony analyses are designed to find, and the unpredictable cladistic behavior of both recent and ancient hybrids due to the different modes of character inheritance and to the possibility of formation of later generation hybrids or introgressants (McDade, 1992; Rieseberg and Ellstrand, 1993). When using molecular characters, knowledge of the genetic control and inheritance of the studied marker may minimize the problem of unpredictable hybrid behavior as compared to morphological characters. However, when sequences are involved, an additional element has to be considered: the distinction between gene trees and species trees due to problems of sampling paralogous instead of orthologous genes (Fitch, 1970; Goodman et al., 1979; Pamilo and Nei, 1988; Patterson, 1988a) and the potential role of lineage sorting (Doyle, 1992). Furthermore, if the sequenced genes belong to a multi-copy family such as the nuclear ribosomal DNA, still one more determinant factor is involved: the degree and speed of intra- and intergenomic homogenization of the repeat types (concerted evolution) before and after hybridization events (Wendel et al., 1995; Brochmann et al., 1996; Baldwin and Robichaux, 1995; Sang et al., 1995; Kim and Jansen, 1994; Campbell et al., 1997). The fact that this gene family may undergo rapid

concerted evolution can simplify sampling and thus be a favorable property for phylogenetic analyses if species are reproductively isolated (Baldwin et al., 1995; Hillis et al., 1991). But, to document reticulate evolution, what in principle would be needed is that either homogenization fails to act across repeat units contributed by different parental species (Baldwin et al., 1995) or that different repeat types are concerted in different hybrids of the same or related progenitors (Wendel et al., 1995). When homogenization fails, whether within a single individual or in different individuals, traces of reticulation can be found in the form of parental molecular markers.

In a previous paper based on one of these multi-copy genes (nuclear ribosomal ITS regions), we presented evidence for extensive reticulation in *Armeria* (Plumbaginaceae) (Fuertes Aguilar et al., 1999b). This followed from a clear geographical pattern obtained from the parsimony analysis of 55 sequences, which was incongruent with patterns of morphological similarity (fig. 1). Each of five major clades in all the 15 optimal trees contained very similar sequences that belonged to several species from the same geographic area. In contrast, sequences from the most intensively sampled taxon (*A. villosa* subsp. *longiaristata*) were spread in three of the five major clades depending on their geographic origin. Our evolutionary interpretation for such a pattern was that extensive gene flow occurs and that biased concerted evolution subsequently tends to homogenize ITS sequences within territories. Alternative explanations for the ITS pattern have been discussed in detail elsewhere (Fuertes Aguilar et al., 1999b). However, none of them accounts for all the evidence.

Our purpose here is (1) to add an input of empirical evidence to our evolutionary interpretation of the ITS data (the behavior of ITS sequences from synthetic hybrids and backcrosses in parsimony analyses), and (2) to discuss its

theoretical implications. We argue that the possibility that a hierarchically structured, and thus cladistically informative, data matrix has resulted from the combination of disturbing forces is theoretically relevant.

EVOLUTIONARY INTERPRETATION FOR ITS DATA

A hypothetico-deductive argument

Before providing the results from the artificial hybrids, in the following paragraphs we examine the reliability of our evolutionary interpretation by (1) defining a hypothetical situation that involves the two elements we suspect occur in *Armeria,* (2) predicting the cladistic patterns expected in a group evolving under such extreme model, (3) and discussing the accuracy of such elements for our study genus (*Armeria*).

Which topology would we expect to find in a cladistic analysis based on ITS data from a group of organisms matching the following two elements?: 1) *Extensive gene flow* (EGF)*,* meaning that hybridization and introgression are very likely to occur whenever different populations, from the same or different species, come into contact and the contacts are frequent enough to allow wide introgression within a territory. 2) *Active and biased concerted evolution* (BCE)*,* meaning that fast non-random homogenization is undergone by ITS regions following crosses between different populations or species.

Under such extreme conditions, we would expect ITS copies to be rapidly homogenized in congeners within territories. In a cladistic context, such a trend would produce a clear pattern: large clades containing very similar sequences, and thus with poor internal resolution, from different species inhabiting the same territory. The bias in homogenization, also called conversional advantage (Li,

1997), would be responsible for sequences becoming increasingly shared within areas. Whether arising from mutations or inherited from ancestors, the 'favored' repeats would tend to become extended via gene flow, while 'non-favored' ITS repeats would tend to be eliminated or confined. Factors that could alter this prediction include high mutation rates within ITS and a high probability that new mutations turned out to be favored over the preexisting ones. Internal resolution in those large clades would be weak, as a consequence of the biased homogenization, unless favored mutations arose in a sufficiently high rate as to introduce additional post-homogenization variability. The sharing of equal or almost equal sequences resulting from high levels of concerted evolution would contribute to low homoplasy levels in a cladistic analysis (Sanderson and Doyle, 1992; see below). Besides the arising of new favored mutations, spatial limitations hindering actual inter-population contact might contribute some variability within territories.

If we compare the cladistic patterns expected from the above model with those obtained from real ITS data (few major clades composed of geographically related taxa, low resolution of terminal and subterminal clades, and low homoplasy levels), the coincidence is obvious (fig. 1; Fuertes Aguilar et al., 1999b). But, what independence evidence is available for the two elements (EGF, BCE) occurring in *Armeria?* Or, how close is our study genus to the model based on evidence external to the wild ITS data.

Ancillary evidence for the EGF-BCE model in Armeria

Independent evidence for the occurrence of EGF in *Armeria* is available, but it is difficult to assess how close to the model is the real situation. For EGF to occur, it requires weak internal isolation barriers. In plants, where it is per se

more feasible than in animals (Grant, 1981), it would be greatly favored if strict allogamy was the breeding system. This assertion is supported by the good correlation between reproductive system and levels of gene flow reported in different groups of vascular plants (Hamrick et al., 1995). In *Armeria,* plants are obligate outcrossers by virtue of an efficient heteromorphic incompatibility system (Baker, 1966). Exceptions include populations from a single species aggregate, *A. maritima* (Northern Pacific, Artic, South-American and populations from heavy metal rich soils in Central Europe). Low internal barriers have been detected in artificial crossings between different species, resulting in as much as 90% of viable pollen in F_1 hybrids (Nieto Feliner et al., 1996; Nieto Feliner, 1997). Besides the cladistic ITS patterns, which are intentionally excluded from the argument at this point to avoid circularity, other evidence suggesting EGF comes from morphological patterns combined with distributional data and ecological preferences, as well as morphometric variation and inferred multigenic control of quantitative characters (Bernis, 1954; Lefèbvre, 1969; Arrigoni, 1970; Philipp, 1974; Stace, 1975; Nieto Feliner, 1987, 1988, 1997; Nieto Feliner et al., 1996).

The fact that among concerted gene families, such as rDNA, new variants are rapidly homogenized and fixed within reproductive groups is well known (Hillis and Davis, 1988; Baldwin et al., 1995). In angiosperms, there is evidence supporting both rapid homogenization (Wendel et al., 1995; Brochmann et al.,1996; Baldwin and Robichaux, 1995) as well as retardation (Sang et al., 1995; Kim and Jansen, 1994; Campbell et al., 1997) the latter sometimes associated with cases where polyploidy and/or apomixis are involved. Experimental evidence for biased concerted evolution in *Armeria*, where most populations are diploid, was demonstrated in artificial F_2 hybrids (Fuertes Aguilar et al., 1999a).

The bias is essential to produce the pattern expected from the hypothetical situation and detected in the real ITS data. If homogenization of the parental ITS types after gene flow events were random, congeners would not all tend to share the same sequence within the same territory, as a consequence, homoplasy would be much greater. However, the cladistic effects of bias depend on the mutation rates. Low mutation rates in ITS are more consistent with the scenario proposed for *Armeria*. A high mutation rate would reduce homogeneity in clades by providing point mutations, which could, if favored by the bias, subsequently produce some terminal and sub-terminal resolution via gene flow and concerted evolution. When sequences of species from isolated regions are compared to those from areas were diversity is high, such as the Iberian Peninsula, divergence between sequences is low (Fuertes Aguilar et al., 1999b; and unpublished data). Since in the former regions variation cannot be increased by interspecific gene flow, this suggests that ITS mutation rates are effectively low in *Armeria.*

Defining models within a pattern-oriented study is questionable (Scotland et al., 1994) but we think there are good reasons to do it here. Hybridization studies aim at inferring something about processes but they must be based on patterns (Arnold, 1997). Because complex patterns may occur when hybridization is involved (Wendel et al., 1995; Rieseberg et al., 1996), it is worth taking over the costs of assuming specific models if it results in an insightful interpretation. Besides, our model does not modify the data set and is not *ad hoc* because its validity can be explored with independent evidence.

In sum, it seems difficult to escape the conclusion that resemblance exists between the model and the real world. EGF and BCE thus appear to be significant, if not essential, factors in producing the ITS pattern.

Cladistic behavior of ITS sequence data from artificial hybrids

Artificial hybrids and backcrosses were raised, between 1989 and 1994, following the procedures in Nieto Feliner et al. (1996). Progenitors and recurrent parents used in the crossing program were *A. colorata* (clade I, fig. 1) and *A. villosa* ssp. *longiaristata* from clade IV (fig. 1). Nuclear ribosomal ITS regions (ITS1+ 5.8S + ITS2) were directly sequenced from PCR products and aligned manually (Fuertes Aguilar et al., 1999a; 1999b). Sequences were deposited in EMBL. The two progenitors of the artificial hybrids differ in six informative positions. Changes in five of those positions support major clades in the wild data set while the sixth supports a smaller clade nested in clade I (fig. 1; Fuertes Aguilar et al., 1999b).

The analyses of ITS sequences presented here are based on data matrices including representative samples spanning all of the variation encountered in the wild as well as a selection from artificial hybrids and backcrosses. The data matrices contain 625 aligned positions, of which 17 were informative in the parsimony analyses. Sequence data from wild taxa used in the analyses were selected from the original matrix described in Fuertes Aguilar et al. (1999b). Identical sequences were removed unless any of the five major clades resolved in that study would be reduced to less than three different taxa as a result of removal. This was done so as to maintain a sample of the geographically based taxonomic composition of the clades. The resulting wild data set contains 33 samples from 23 different taxa (Fig. 1).

Sequences from artificial hybrids in Fuertes Aguilar et al. (1999a) were selected for parsimony analysis in this study. From each of the artificial hybrid generations (F_1, F_2, $B_1 \rightarrow$ *colorata*, $B_1 \rightarrow$ *villosa longiaristata*, $B_2 \rightarrow$ *villosa longiaristata*), identical sequences were removed for the present study. As a result, of the 30 samples from artificial hybrids available, only 11 were used here (Appendix 1).

Analyses were performed on four data sets. All contain the same wild subset of 33 terminals taken from Fuertes Aguilar et al. (1999b) but each of the four was completed with a different selection of the artificial hybrid sequences. The four data sets comprised: 1) wild data plus all the hybrids and backcrosses (B_1, B_2), 2) wild data plus the F_1 generation, 3) wild data plus the backcrossed generations (B_1, B_2), 4) wild data plus the F_2 generation.

Parsimony analyses were undertaken using PAUP* 4.0b2 (Swofford, 1999). Character-states were treated as unordered and polymorphisms, as well as gaps resulting from the alignment, were treated as missing data. Options used were MULPARS and TBR, for tree searching, and ACCTRAN, for character optimization. Ten replicate heuristic searches with random taxon addition were performed to maximize the probability of hitting all islands of most parsimonious trees (Maddison, 1991). However, to circumvent the collapsing of the computer memory in the analysis of data set 1 (see below), for this matrix we used an alternative strategy: 1000 replicates of random taxon addition limiting the number of trees in each replicate to 100. Bootstrap values were computed using the fast stepwise addition option with 100,000 replicates. Previously, we had compared bootstrap values for two of the matrices mentioned above with those obtained using 100 replicates each with 10 random taxon addition, and found minor differences. This is in accordance with Mort et al. (2000). Cladograms were

rooted using *Psylliostachys suworowii,* following a phylogeny of Plumbaginaceae based on *rbcL* sequence data (Lledó et al., 1998).

The sequences of the artificial hybrids and backcrosses do not contain new character-states as compared to the wild sequences. They differ only by presenting additive polymorphic sites in some of the six positions in which the selected progenitors differ, or in one of the two parental states. Since the data come from direct sequencing, polymorphisms in artificial hybrids are probably the result of co-occurrence of ITS repeats from both parents in the hybrids. However, most of those polymorphic sites show homogenization in the F_2, presumably due to concerted evolution (Fuertes Aguilar et al., 1999a). Apart from the polymorphic sites resulting from hybridization, only two other non-informative polymorphisms are found in the artificial hybrids.

Polymorphic characters introduced in the data set with the F_1 artificial hybrids are expected to produce a certain amount of disruption or uncertainty as compared to the analysis of the wild data. Homogenization in some of those characters in later generation hybrids is likely to improve the results. The purpose of these analyses is to explore the effect that homogenization of polymorphic sites produces in the topology of the combined (wild + artificial) data sets and in the placement of hybrid terminals. If homogenization in the F_2 reduces both topological disruption and differences with topology of the wild data set, we may think that some of the wild sequences have resulted from the same homogenizing mechanism. To this end, we have presented the analyses from the most disturbed to the least disturbed (table 1, Fig. 2), i.e. from the analyses containing hybrids with more polymorphic sites to the analyses containing hybrids with more homogenized sites. The consistency and retention indices are virtually

the same in the different analyses (CI = 0.82, RI = 0.95-0.96, table 1) as are the number of steps in the most parsimonious reconstructions (L= 103).

The analysis of the wild subset of the data plus all the hybrids and backcrosses (data set 1) found 17405 most parsimonious trees, the strict consensus recognizing only three of the six major geographically-based clades (II, III and V). The combinable component (semistrict) consensus of those fundamental trees recovers one more major clade (IV) as well as a geographically-based smaller clade (Ia) nested within I and where one of the progenitors belongs (*A. colorata*). Analyzing the matrix that contains the wild subset of the data plus the F_1 progeny (data set 2), we obtained the same results in terms of wild major clades recovered (table 1). When the matrix including the wild samples and the backcrossed progeny is run (data set 3), 1404 most parsimonious cladograms are obtained. The number of major clades recovered in the strict and combinable components consensus is the same as in the two above data sets. The fourth analysis (data set 4), which includes the F_2 samples whose polymorphisms are homogenized to a greater extent, yields similar results to those of the latter. But the combinable components consensus contains most of the resolution found in the cladogram of the wild data set.

Along these analyses, the degree of homogenization in the six polymorphic sites in which the two parental taxa differ goes in parallel both with the reduction in number of most parsimonious trees as well as with increasing resolution and a more distal placement of hybrid terminals. In poorly resolved consensus cladograms artificial hybrids, at least the ones with more polymorphisms, are placed at or near the base. As the tree topology improves, in sequences with some or high degree of homogenization, artificial hybrids and backcrosses tend to be placed in or near clades containing the parents.

Expectedly, as artificial hybrids move distally, polytomies tend to do so. Departures from this trend are partly due to the fact that we have included in the analyses all the different artificial hybrid sequences detected independently of whether or not they are quantitatively representative of the variation found.

Possible results in the cladistic analysis of the combined (artificial + wild) data sets as compared to the 'wild' data set that would question the likelihood of our interpretation include the following: very different placement of hybrids terminals from the same offspring in the cladogram, a significant increase in homoplasy, or the placement of hybrids in clades that do not exist in the natural data set. However, none of these things happened. Instead, proven hybrids, with some degree of homogenization fall within the major clades recovered from the analysis of the wild taxa. Neither does consistency index decrease significantly in analyses with hybrid sequences (0.82 vs. 0.84 in wild data set in Fuertes Aguilar et al., 1999b). If the behavior of these artificial hybrids is representative of products of natural hybridization, our results suggest that hybridization and introgression in the wild results in some of the natural ITS clades containing an increasing number of terminals, and this is consistent with the ITS pattern found in the wild data set. As expected, these results contrast with McDade's (1992) study of the impact of F_1 hybrids in cladistic analyses, the primary reason being that morphological characters, on which her analysis is based, lack homogenization mechanisms and are subject to different modes of inheritance.

Because polymorphisms have been treated as missing data, we may think that the lack of conflicts caused by artificial hybrids in the analyses is somewhat optimistic. However, in the analysis of the wild taxa, there are no polymorphic states in the 12 nucleotide positions supporting the major clades, with a single exception in one terminal for one of the characters (Fuertes Aguilar et al.,

1999b). This would suggest that invariant positions for natural clades have been effectively homogenized and fixed. Besides, an analysis of data set 2 using an alternative treatment of the polymorphic states, under *polymorphism parsimony* (Farris, 1978; Felsenstein, 1979) gave the same resolution and number of trees although, expectedly, the length increased to 169.

Artificial hybrids provide relevant information for assessing the feasibility of the scenario we propose for ITS in *Armeria*. Since they somehow accomplish the two elements proposed (they are true hybridization products and they have undergone some degree of homogenization in their ITS sequences), their performance in cladistic analysis gives additional support to our interpretation.

Homoplasy levels

Another aspect from the results of the wild ITS data that is consistent with expectations from the EGF-BCE model is the low homoplasy levels. There are several possible sources of homoplasy, some of them resulting from independent origins of the same character state (a given nucleotide in a given aligned position), and some of them caused by mixing different historical signals (Doyle, 1996). Here, we face the opposite: low homoplasy levels even if one of the potential sources of mixed signal homoplasy, hybridization, is apparently involved. The retention and consistency indices of the *Armeria* wild data set (Fuertes Aguilar et al., 1999b) are high [CI=0.84 excluding uninformative characters, RI=0.96] as compared to two recent literature surveys (Sanderson and Donoghue, 1996; Givnish and Sytsma, 1997). But the highest CIs from sequence data in the same survey also correspond to ribosomal DNA. Therefore, it is apparent that low homoplasy levels are the rule when analyzing nrDNA sequences and that homogenization and fixation of repeat types within

reproductive groups may contribute to it (Hillis and Davis, 1988). In fact, simulation studies suggest that high levels of concerted evolution are not problematic when reconstructing organismal phylogenies, not gene histories, and that such high levels may be realistic for nrDNA (Sanderson and Doyle, 1992). The same authors conclude that, above an intermediate level, as concerted evolution levels increase homoplasy decreases. Therefore, high levels of concerted evolution do explain low homoplasy levels. However, such levels of concerted evolution cannot, by themselves, explain the *Armeria* ITS data. Concerted evolution homogenizes multigene copies among reproductive groups. In predominantly divergent evolutionary scenarios sequences are homogenized within species. But in our data some sequences are shared by conspecific terminals, others by non-conspecifics (Fuertes Aguilar et al., 1999b). Therefore, for concerted evolution to be responsible for such sequences shared by different species, gene flow among them has to have occurred. The fact that those non-conspecific terminals sharing the same sequence are geographically related only adds strength to the whole argument. In sum, concerted evolution per se does not explain our patterns but it does if EGF is involved.

However, other aspects of our interpretation of ITS data are not so clear. The main reason adduced by Skála and Zrzavý (1994) for treating terminal taxa equally in cladistic analysis independently of their possible hybrid origin is that there are no “tools for deductively distinguishing between various kinds of character conflict”. We have minimum character conflict in our data set. However, individual contribution by different forces potentially involved is unclear. Specifically, if concerted evolution is promoting homogenization of ‘favored’ ITS sequences among species and within territories: What is the role of backcrosses in the homogenization? From artificial hybrids we know that two generations may

be enough to gain some homogenization in the different positions (Fuertes Aguilar et al., 1999a) but this process might be either enhanced or counteracted by backcrossing towards one of the parents. When backcrossing towards the parent possessing the 'non-favored' nucleotides occur, the bias effect is neutralized to some extent. We found this effect in our artificial B_1 and B_2 towards *A. villosa* subsp. *longiaristata*. Therefore, even if the two elements, EGF and BCE, are determinant in *Armeria*, other forces may have an influence too, maybe in explaining slight departures from the cladistic predictions. Due to these uncertainties, the degree of EGF demanded is unclear too.

CLADISTIC IMPLICATIONS

Following the previous discussion, we may conclude that there is evidence that EGF occurs and concerted evolution operates actively in *Armeria*, simply because it is our best explanation for all the independent available data. However, the question may arise if a cladistic analysis makes sense when the hierarchical structure that we are looking for does not result from a predominantly divergent scenario. It is true that standard cladistic analysis aims to reveal hierarchical groups by searching for a pattern of character-distribution (nested sets of synapomorphies) and that reticulation may distort such a pattern (McDade, 1995; Rieseberg and Ellstrand, 1993). However, in a more neutral view, cladograms can also be considered as summaries of information regarding characters (Nelson and Platnick, 1981; Skála and Zrzavý, 1994). It is under this approach, that we analyzed our ITS data set.

Hierarchical structure of the data

Even if there is justification to analyze our data set cladistically, we might fail to find informative, recoverable signals. Does our data set contain a hierarchical structure? The consistency index is an appropriate test for hierarchy, but other qualities of the data may be explored that are independent of the amount of homoplasy (Goloboff, 1991). Data decisiveness (DD) measures the quality of the data by the extent to which it allows to chose some cladograms over others.

To calculate data decisiveness (DD; Goloboff, 1991) for the wild data set in Fuertes Aguilar et al. (1999b), an approximation to $\bar{S}$, the mean length of all possible cladograms, was computed. This was done from the mean of 100,000 randomly resolved trees generated by PAUP. In our wild data set, DD is high (0.92) implying that the data is decisive. Another parameter that has been used to measure the phylogenetic signal of the data is skewness in distribution of tree lengths, DTL (Hillis, 1991). This was calculated from the same 100,000 trees generated by PAUP. The value obtained (g_1= –0.33, $p<0.01$) is significantly more skewed than expected from random data (Hillis and Huelsenbeck, 1992). This is so despite the fact that characters in the data matrix divide terminals into groups of similar size, a property that makes the DTL more symmetrical and thus inappropriate in certain circumstances (Källersjö et al.,1992). In other words, DTL indicates hierarchical structure in the data set even though the distribution of character-states may be reducing the skewness and thus causing a more conservative measure of the hierarchy through this test. Therefore, both parameters are consistent in suggesting that the wild data matrix does contain a significant cladistic structure.

What signal are we recovering?

But if such structure has arisen partly from reticulation and homogenization of ITS sequences, another question would be what are we really analyzing? We do not need to assume the processes responsible for character distribution to do cladistics (Patterson, 1988b). In *Armeria*, however, independent evidence has been used *a posteriori* to help explain cladistic patterns, and such patterns are very difficult to interpret unless reticulation is invoked. Thus, once concluding that real world in *Armeria* is close to the model, it seems inescapable to ask what do the terminals in the same clade share?

Vrana and Wheeler (1992) state that trees of multi-copy concerted nuclear genes may yield a pattern that is not directly relevant to the question of hierarchy among the organisms. We think that our ITS data, partly a result of concerted evolution, is in fact providing relevant information but it might be true that it is not too informative about hierarchy resulting from common ancestry. In other words, because biased concerted evolution homogenizes sequences among reproductive groups, when the reproductive groups involve different ‘species’, information on evolutionary history individual species is being partially erased. However, this process is leaving traces of gene flow in such a way that it appears to be telling us something else. It could have happened that gene flow occurred but the traces left were obscure. However, this is not the case as revealed by the clear geographical structure of the data. That the same sequences are linked to similar geographical areas rather than to species implies that the pattern has evolutionary implications at the organismal level and not just at the molecular level. Such geographical structure is so apparent that the gene cladogram can be transformed without any conflict into an area cladogram, by recognizing the major clades as terminals and labeling them with the area where the involved

sequence is distributed. However, the evidence that geographic components arise not from vicariance but from two potentially conflicting elements combined, such as hybridization and concerted evolution of multi-copy genes, make our results worth of attention, especially for those who study groups that involve reticulation using molecular data. In plants, the ITS region is the most widely used molecular marker in phylogenetic studies below the family level. Ultimately, to the question of what do some of the terminals share, we could tentatively answer a recent hybridization event with an individual carrying a given piece of DNA, which is predominant in a territory.

ACKNOWLEDGEMENTS

We are grateful to Bruce Baldwin, Paolo Caputo, Javier Francisco-Ortega and Chris Humphries for their helpful comments to an earlier version of this manuscript, also to Gonzalo Giribet and Jyrki Muona for constructive criticisms and suggestions to a later version. This work has been supported by grants DGICYT PB94-0110 and DGES PB97-1146 of the Spanish Dirección General de Enseñanza Superior e Investigación Científica.

REFERENCES

Arnold, M. L. (1997). “Natural Hybridization and Evolution”. Oxford University Press, New York.

Arrigoni, P. V. (1970). Contributto alla conoscenza delle Armerie sardo-corse. *Webbia* **25**, 137-182.

Baker, H. G. (1966). The evolution, functioning and breakdown of heteromorphic incompatibility systems, I. The Plumbaginaceae. *Evolution* **20**, 349-368.

Baldwin, B. G., and Robichaux, R. H. (1995). Historical biogeography and ecology of the Hawaiian silversword alliance (Asteraceae). *In* "Hawaiian biogeography: evolution on a hot-spot archipelago" (W.L. Wagner, and V. A. Funk, Eds.), pp 259-287. Smithsonian Institution Press, Washington DC.

Baldwin, B. G., Sanderson, M. J., Porter, J. M., Wojciechowski, M. F., Campbell, C. S., and Donoghue, M.J. (1995). The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. *Ann. Mo. Bot. Gard.* **82**, 247-277.

Bernis, F. (1954). Revisión del género *Armeria* Willd. con especial referencia a los grupos ibéricos. Parte segunda (descriptiva de los grupos ibéricos). *An. Inst. Bot. A. J. Cavanilles* **11(2)**, 5-288.

Brochmann, C., Nilsson, T. and Gabrielsen, T. M. (1996). A classic example of postglacial allopolyploid speciation re-examined using RAPD markers and nucleotide sequences: *Saxifraga osloensis* (Saxifragaceae). *Symb. Bot. Ups.* **31**(3), 75-89.

Campbell, C. S., Wojciechowski, M. F., Baldwin, B. G., Alice, L. A., and Donoghue, M. J. (1997). Persistent nuclear ribosomal DNA sequence polymorphism in the *Amelanchier* agamic complex (Rosaceae). *Mol. Biol. Evol.* **14**, 81-90.

Doyle, J. J. (1992). Gene trees and species trees: molecular systematics as one-character taxonomy. *Syst. Bot.* **17**, 144-163.

Doyle, J. J. (1996). Homoplasy connections and disconnections: genes and species, molecules and morphology. *In* "Homoplasy, The recurrence of similarity in Evolution" (M. J. Sanderson, and L. Hufford, Eds), pp. 67-89. Academic Press, San Diego.

Farris, J. S. (1978). Inferring phylogenetic trees from chromosome inversion data. *Syst. Zool.* **27**, 275-284.

Felsenstein, J. (1979). Alternative methods of phylogenetic inference and their interrelationship. *Syst. Zool.* **28**, 49-62.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99-113.

Fuertes Aguilar, J., Rosselló, J. A., and Nieto Feliner, G. (1999a). Nuclear ribosomal DNA (nrDNA) concerted evolution in natural and artificial hybrids of *Armeria* (Plumbaginaceae). *Mol. Ecol.* **8**, 1341-1346.

Fuertes Aguilar, J., Rosselló, J. A., and Nieto Feliner, G. (1999b). Molecular evidence for the compilospecies model of reticulate evolution in *Armeria (Plumbaginaceae). Syst. Biol.* **44**, 735-754.

Funk, V. A. (1985). Phylogenetic patterns and hybridization. *Ann. Misssouri Bot. Gard.* **72**, 681-715.

Givnish, T. J., and Sytsma, J. (1997). Homoplasy in molecular vs. morphological data: the likelihood of correct phylogenetic inference. *In* "Molecular evolution and adaptive radiation" (T.J Givnish, and J. Sytsma, Eds), pp. 55-101. Cambridge University Press, Cambridge.

Goloboff, P. A. (1991). Homoplasy and the choice among cladograms. *Cladistics* **7**, 215-232.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., and Matsuda, G. (1979). Fitting the gene lineage into its species lineage: A parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.* **28**, 132-168.

Grant, V. (1981). "Plant Speciation", 2nd ed. Columbia University Press, New York.

Hamrick, J. L., Godt, M. J. W., and Sherman-Broyles, S. L. (1995). Gene flow among plant populations: evidence from genetic markers. *In* "Experimental and molecular approaches to plant biosystematics" (P. E. Hoch, and A. G. Stephenson Eds), pp. 215-232. Missouri Botanical Garden, St. Louis, MO.

Hillis, D. M. (1991). Discriminating between phylogenetic signal and random noise in DNA sequences. *In* "Phylogenetic analysis of DNA sequences" (M. M. Miyamoto, and J. Cracraft, Eds), pp. 278-294. Oxford University Press, Oxford.

Hillis, D. M., and Davis, S. K. (1988). Ribosomal DNA: intraspecific polymorphism, concerted evolution, and phylogeny reconstruction. *Syst. Zool.* **37**, 63-66.

Hillis, D.M., and Huelsenbeck, J.P. (1992). Signal, noise, and reliability in molecular phylogenetic analyses. *J. Hered*. **83,** 189-195.

Hillis, D. M., Moritz, C., Porter, C. A., and Baker, R. J. (1991). Evidence for biased gene conversion in concerted evolution of ribosomal DNA. *Science* **251**, 308-310.

Humphries, C. J. (1983). Primary data in hybrid analysis. *In* "Advances in Cladistics 2". (N. I. Platnick, and V. A. Funk, Eds), p.p. 89-103. Columbia University Press, N.York.

Källersjö, M., Farris, J. S., Kluge, A. G., and Bult, C. (1992). Skewness and permutation. *Cladistics* **8**, 275-287.

Kim, K.-J., and Jansen, R. K. (1994). Comparisons of phylogenetic hypothesis among different data sets in dwarf dandelions (*Krigia*): additional information from internal transcribed spacers sequences of nuclear DNA. *Plant Syst. Evol.* **190**, 157-185.

Lefèbvre, C. (1969). Populations d'*Armeria* Willd. le long de la Mer Baltique et de la Mer du Nord. *Bull. Soc. Roy. Bot. Belgique* **102**, 293-303.

Li, W.-H. (1997). "Molecular Evolution". Sinauer Ass., Sunderland, MA.

Lledó, M. D., Crespo, M. B., Cameron, K. M., Fay, M. F., and Chase, M. W. (1998). Systematics of Plumbaginaceae based upon cladistic analysis of rbcL sequence data. *Syst. Bot.* **23**, 21-29.

Maddison, D. R. (1991). The discovery and importance of multiple islands of most-parsimonious trees. *Syst. Zool*. **40**, 315-328.

McDade, L. A. (1992). Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. *Evolution* **46**, 1329-1346.

McDade, L. A. (1995). Hybridization and Phylogenetics. *In* "Experimental and molecular approaches to plant biosystematics" (P. C. Hoch, and A. G. Stephenson, Eds), pp. 305-331. Missouri Botanical Garden, St. Louis, MO).

Mort, M. E., Soltis, P. S., Soltis, D. E., and Mabry, M. L. (2000). Comparison of three methods for estimating internal support on phylogenetic trees. *Syst. Biol.* **49**, 160-171.

Nelson, G., and Platnick, N. (1981). Systematics and Biogeography, Cladistics and Vicariance. Columbia University Press, N.York.

Nieto Feliner, G. (1987). El género *Armeria* (Plumbaginaceae*)* en la Península Ibérica: aclaraciones y novedades para una síntesis. *An. Jard. Bot. Madrid* **44**, 319-348.

Nieto Feliner, G. (1988). Flujo génico en *Armeria* (Plumbaginaceae*)* en la Península Ibérica. *Lagascalia* **15** (extra), 233-236.

Nieto Feliner, G. (1997). Natural and experimental hybridization in *Armeria* (Plumbaginaceae): *A. salmantica*. *Int. J. Plant Sci.* **158**, 585-592.

Nieto Feliner, G., Izuzquiza, A., and Lansac, A.R. (1996). Natural and experimental hybridization in *Armeria* (Plumbaginaceae): *A. villosa* subsp. *carratracensis*. *Plant Syst. Evol.* **201**, 163-177.

Pamilo, P., and Nei, M. (1988). Relationships between gene trees and species trees. *Mol. Biol. Evol*. **5**, 568-583.

Patterson, C. (1988a). Homology in classical and molecular biology. *Mol. Biol. Evol.* **5**, 603-625.

Patterson, C. (1988b). The impact of evolutionary theories on systematics. *In* "Prospects in Systematics" (D. L. Hawksworth, Ed), p.p. 59-91. Clarendon Press, Oxford.

Philipp, M. (1974). Morphological and genetical studies in the *Armeria* maritima Aggregate. *Bot. Tidsskr.* **69**, 40-51.

Rieseberg, L. H., and Ellstrand, N. C. (1993). What can molecular and morphological markers tell us about Plant Hybridization? *Crit. Rev. Plant Sci.* **12**, 213-241.

Rieseberg L. H., and Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evol. Trends Plants* **5,** 65-84.

Rieseberg, L.H., Whitton, J., and Linder, C.R. (1996). Molecular marker incongruence in plant hybrid zones and phylogenetic trees. *Acta Bot. Neerl*. **45**, 243-262.

Sanderson, M. J., and Donoghue, M. J. (1996). The relationship between homoplasy and confidence in a phylogenetic tree. *In "*Homoplasy, The recurrence of similarity in Evolution" (M. J. Sanderson, and L. Hufford, Eds), pp. 67-89. Academic Press, San Diego.

Sanderson, M. J., and Doyle, J. J. (1992). Reconstruction of organismal and gene phylogenies from data on multigene families: concerted evolution, homoplasy, and confidence. *Syst. Biol.* **41**, 4-17.

Sang, T., Crawford, D. J., and Stuessy, T. F. (1995). Documentation of reticulate evolution in peonies (*Paeonia*) using internal transcribed spacer sequences of nuclear ribosomal DNA: implications for biogeography and concerted evolution. *Proc. Natl. Acad. Sci. USA* **92**, 6813-6817.

Scotland, R. W., Siebert, D. J., and Williams, D. M. (Eds) (1994). "Models in Phylogeny reconstruction". Clarendon Press, Oxford.

Skála, Z., and Zrzavý, J. (1994). Phylogenetic reticulations and cladistics: discussion of methodological concepts. *Cladistics* **10**, 305-313.

Stace, C. A. (1975). "Hybridization and the Flora of the British Isles". Academic Press, London.

Swofford, D. L. (1999). PAUP*, Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, Massachusetts.

Vrana, P., and Wheeler, W. (1992). Individual organisms as terminal entities: laying the species problem to rest. *Cladistics* **8**, 67-72.

Wendel, J. F., Schnabel, A., and Seelanan, T. (1995). Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proc. Natl. Acad. Sci. USA* **92**, 280-284.

Legends:

Fig. 1.— Analysis of nuclear ribosomal DNA ITS1+ 5.8S + ITS2 sequences from wild samples of *Armeria* in the Iberian peninsula. (A) Combinable components (semistrict) consensus cladogram based on a subset of the wild samples used in Fuertes Aguilar et al. (1999b). Shaded rectangles and corresponding roman numerals indicate major

clades, each containing sequences from three or more different taxa and corresponding to a distinct geographical area. (B) Outline of the ITS geographical pattern showing the correspondence between major clades and geographical areas. Accession numbers of the sequences deposited in EMBL are (order follows that in the cladogram): AJ225597, AJ225586, AJ225593, AJ225587, AJ225564, AJ225565, AJ225572, AJ225563, AJ225571, AJ225572, AJ225596, AJ225599, AJ225578, AJ225577, AJ225601, AJ225582, AJ225602, AJ225603, AJ225591, AJ225610, AJ225584, AJ225581, AJ225579, AJ225600, AJ225583, AJ225586, AJ225611, AJ225606, AJ225608, AJ225612, AJ225574, AJ225590, AJ225575, AJ132446.

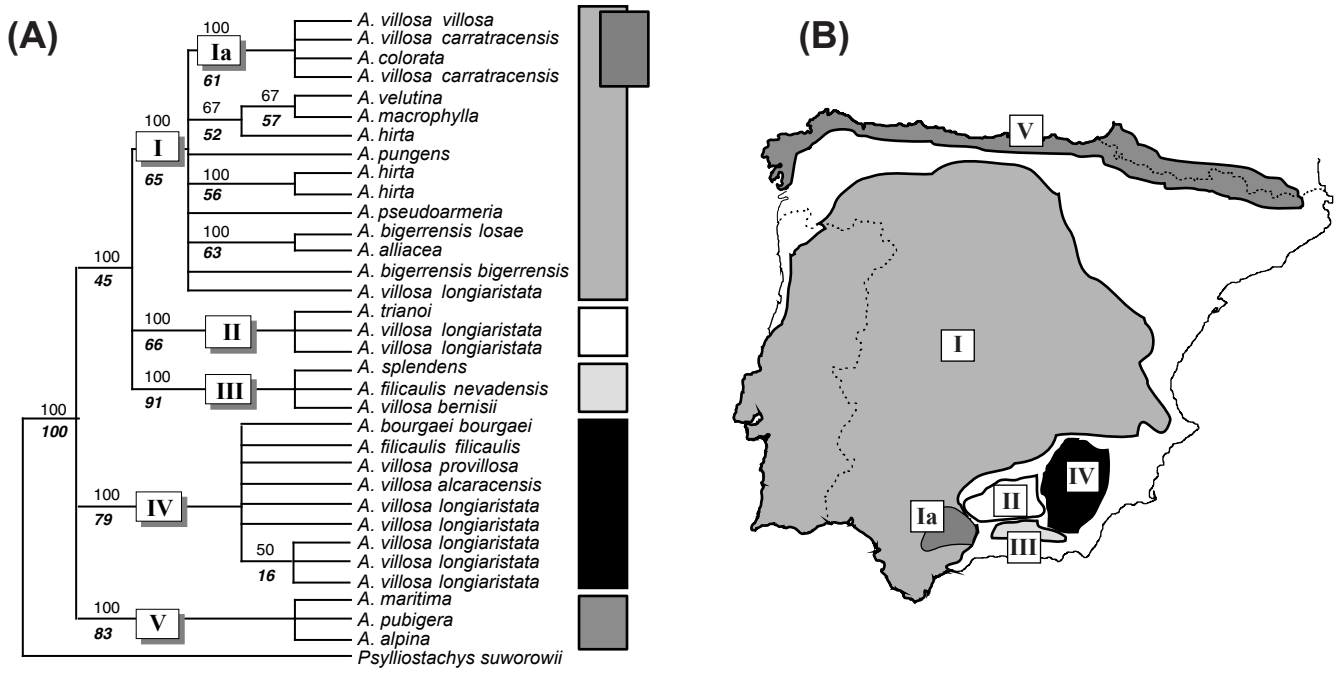
Fig. 2.– Parsimony analysis of nuclear ribosomal DNA ITS data sets combining sequences from wild samples with artificial hybrids and backcrosses obtained by crossing *Armeria colorata* and *A. villosa* subsp. *longiaristata*. Combinable components (semistrict) consensus cladograms from analyses of four combined matrices: a) wild samples + F_1, F_2, and backcrosses (B_1, B_2); b) wild samples + F_1; c) wild samples + backcrosses (B_1, B_2); d) wild samples + F_2. Numbers above nodes indicate the percentage of fundamental cladograms where the clade appears. Italicized numbers below nodes indicate bootstrap values. Roman numerals indicate major clades from the analysis of wild data (see fig.1).

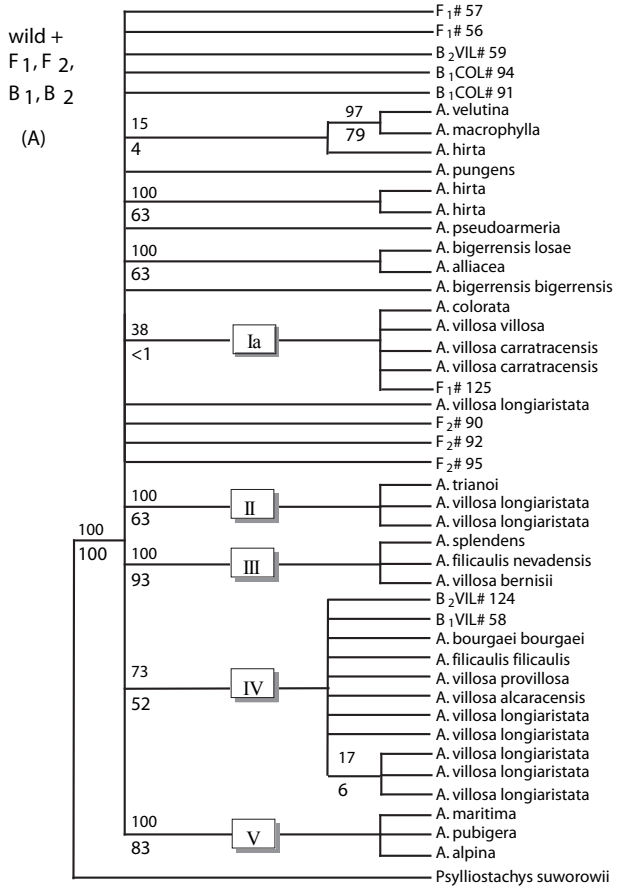
Table 1.-- Cladistic behavior of nuclear ribosomal DNA ITS1+ 5.8S + ITS2 sequences from artificial hybrids of *Armeria* in analyses performed on matrices constructed by combining wild sequences with those from different generations of artificial hybrids.

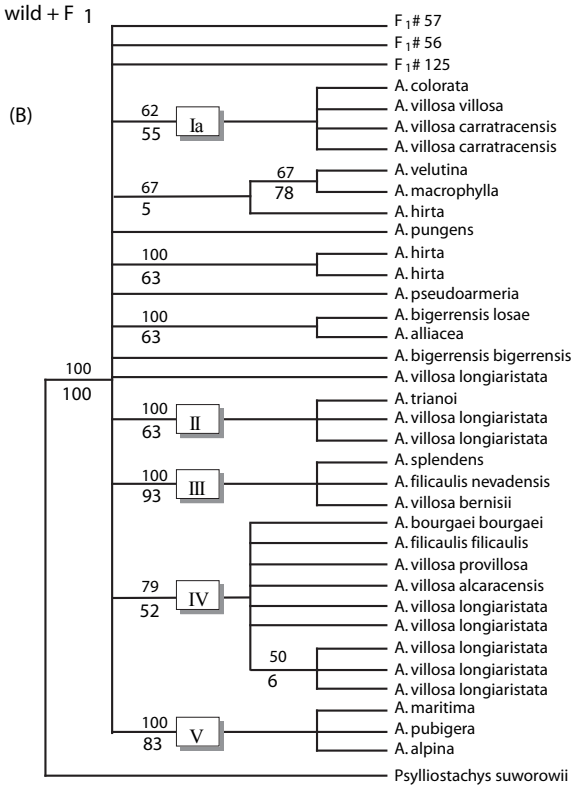
Table 1

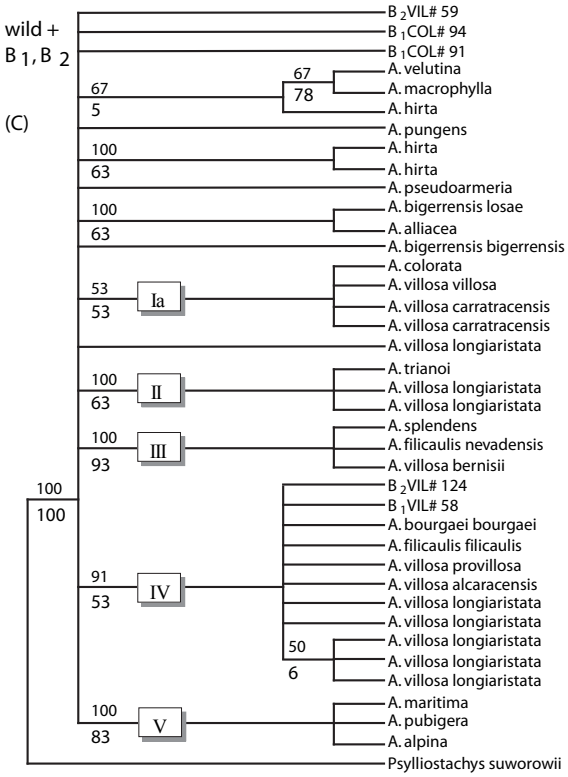
	Number of most parsimonious cladograms	C.I.	R.I.	Number of steps	Major clades recovered in		Placement of artificial hybrids in	
					Strict consensus*	Combinable components consensus*	Strict consensus	Combinable components consensus
Wild data set + all hybrids and backcrosses	17405	0.82	0.96	103	II, III, V	Ia, II, III, IV, V	- base of the ingroup	- base of the ingroup - clade IV
Wild data set + F_1 hybrids	2652	0.82	0.95	103	II, III, V	Ia, II, III, IV, V	- base of the ingroup	- base of the ingroup
Wild data set + backcrosses (B_1 and B_2)	1404	0.82	0.95	103	II, III, V	Ia, II, III, IV, V	- base of the ingroup	- base of the ingroup - clade IV
Wild data set + F_2 hybrids	612	0.82	0.95	103	I, II, III, IV, V	I, Ia, II, III, IV, V	- clade I	- sister to clade Ia within clade I

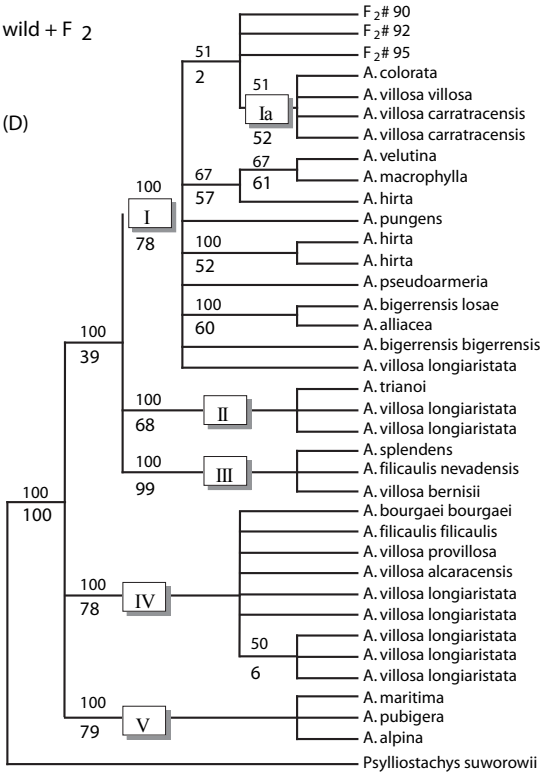
* - Wild clades which, besides the wild terminals, include also artificial hybrids are also considered recovered.


(A)
(B)
A. villosa villosa
A. villosa carratracensis
A. colorata
A. villosa carratracensis
A. velutina
A. macrophylla
A. hirta
A. pungens
A. hirta
A. hirta
A. pseudoarmeria
A. bigerrensis losae
A. alliacea
A. bigerrensis bigerrensis
A. villosa longiaristata
A. trianoi
A. villosa longiaristata
A. villosa longiaristata
A. splendens
A. filicaulis nevadensis
A. villosa bernisii
A. bourgaei bourgaei
A. filicaulis filicaulis
A. villosa provillosa
A. villosa alcaracensis
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. maritima
A. pubigera
A. alpina
Psylliostachys suworowii
Ia
I
II
III
IV
V


wild +
F 1, F 2,
B 1, B 2
(A)
F 1# 57
F 1# 56
B 2VIL# 59
B 1COL# 94
B 1COL# 91
A. velutina
A. macrophylla
A. hirta
A. pungens
A. hirta
A. hirta
A. pseudoarmeria
A. bigerrensis losae
A. alliacea
A. bigerrensis bigerrensis
A. colorata
A. villosa villosa
A. villosa carratracensis
A. villosa carratracensis
F 1# 125
A. villosa longiaristata
F 2# 90
F 2# 92
F 2# 95
A. trianoi
A. villosa longiaristata
A. villosa longiaristata
A. splendens
A. filicaulis nevadensis
A. villosa bernisii
B 2VIL# 124
B 1VIL# 58
A. bourgaei bourgaei
A. filicaulis filicaulis
A. villosa provillosa
A. villosa alcaracensis
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. maritima
A. pubigera
A. alpina
Psylliostachys suworowii
Ia
II
III
IV
V
15
4
97
79
100
63
100
63
38
<1
100
63
100
100
73
52
17
6
100
83


wild + F 1
(B)
F 1# 57
F 1# 56
F 1# 125
A. colorata
A. villosa villosa
A. villosa carratracensis
A. villosa carratracensis
A. velutina
A. macrophylla
A. hirta
A. pungens
A. hirta
A. hirta
A. pseudoarmeria
A. bigerrensis losae
A. alliacea
A. bigerrensis bigerrensis
A. villosa longiaristata
A. trianoi
A. villosa longiaristata
A. villosa longiaristata
A. splendens
A. filicaulis nevadensis
A. villosa bernisii
A. bourgaei bourgaei
A. filicaulis filicaulis
A. villosa provillosa
A. villosa alcaracensis
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. maritima
A. pubigera
A. alpina
Psylliostachys suworowii
Ia
II
III
IV
V
62
55
67
78
67
5
100
63
100
63
100
100
100
63
100
93
79
52
50
6
100
83


wild +
B 1, B 2
(C)
B 2VIL# 59
B 1COL# 94
B 1COL# 91
A. velutina
A. macrophylla
A. hirta
A. pungens
A. hirta
A. hirta
A. pseudoarmeria
A. bigerrensis losae
A. alliacea
A. bigerrensis bigerrensis
A. colorata
A. villosa villosa
A. villosa carratracensis
A. villosa carratracensis
A. villosa longiaristata
A. trianoi
A. villosa longiaristata
A. villosa longiaristata
A. splendens
A. filicaulis nevadensis
A. villosa bernisii
B 2VIL# 124
B 1VIL# 58
A. bourgaei bourgaei
A. filicaulis filicaulis
A. villosa provillosa
A. villosa alcaracensis
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. maritima
A. pubigera
A. alpina
Psylliostachys suworowii
Ia
II
III
IV
V
67
5
67
78
100
63
100
63
53
53
100
63
100
93
100
100
91
53
50
6
100
83


wild + F 2
(D)
F 2# 90
F 2# 92
F 2# 95
A. colorata
A. villosa villosa
A. villosa carratracensis
A. villosa carratracensis
A. velutina
A. macrophylla
A. hirta
A. pungens
A. hirta
A. hirta
A. pseudoarmeria
A. bigerrensis losae
A. alliacea
A. bigerrensis bigerrensis
A. villosa longiaristata
A. trianoi
A. villosa longiaristata
A. villosa longiaristata
A. splendens
A. filicaulis nevadensis
A. villosa bernisii
A. bourgaei bourgaei
A. filicaulis filicaulis
A. villosa provillosa
A. villosa alcaracensis
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. villosa longiaristata
A. maritima
A. pubigera
A. alpina
Psylliostachys suworowii
I
Ia
II
III
IV
V
51
2
51
52
67
61
67
57
100
78
100
52
100
60
100
39
100
68
100
99
100
100
100
78
50
6
100
79